DataFirst

DataFirst Technical Papers

Recalibrating the OHSs to adjust for sampling changes

by
Takwanisa Machemedze, Andrew Kerr
and
Martin Wittenberg

Technical Paper Series
Number 28

About the Author(s) and Acknowledgments

Takwanisa Machemedze - Data Analyst, DataFirst, University of Cape Town
Andrew Ker - Senior Researcher, DataFirst, University of Cape Town
Martin Wittenberg - Director, DataFirst and Professor, School of Economics, University of Cape Town

Recommended citation

Machemedze, T., Kerr, A., Wittenberg, M., (2014). Recalibrating the OHSs to adjust for sampling changes. A DataFirst Technical Paper 28. Cape Town: DataFirst, University of Cape Town

---

DataFirst, University of Cape Town, Private Bag, Rondebosch, 7701, Tel: (021) 650 5708,
Email: info@data1st.org/support@data1st.org

DataFirst

Recalibrating the OHSs to adjust for sampling changes

Takwanisa Machemedze, Andrew Kerr and Martin Wittenberg
DataFirst Technical Paper 28
University of Cape Town
31 October 2014

Introduction

Twenty years after the end of Apartheid South Africa has much more information about the living conditions of its population than it did at the beginning. Nevertheless many of the important questions about the trajectory of South Africa's development require a coherent picture of the starting point. As Branson and Wittenberg (2007) note the 1995 October Household Survey (OHS) has frequently been used as the anchor for post-Apartheid comparisons. They point out, however, that on many dimensions the 1995 OHS looks anomalous when compared to other surveys conducted in the early post-Apartheid period (see also Wittenberg, forthcoming). This of course raises the question whether there is some other suitable base.

Kerr and Wittenberg (2013) have suggested that there is an important discontinuity between **all** the early OHSs (pre-1999) and the later surveys. They suggest that the sampling design of the OHSs led to the systematic under-sampling of small households. The problem that they identified can be shown in Figure 1. We see a doubling in the number of one person households measured according to the household surveys between October 1996 and February 2000, with most of the increase happening after October 1998.

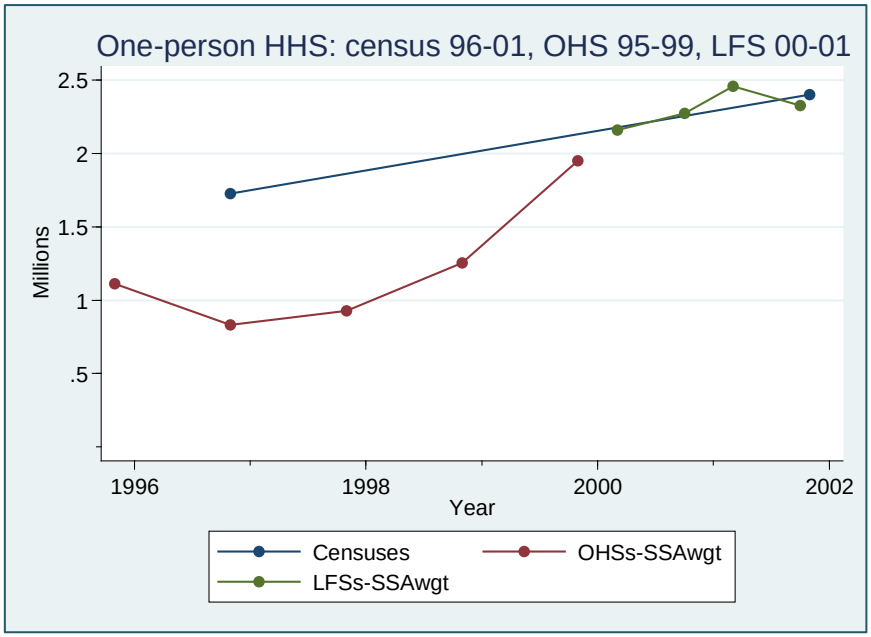

Figure 1 One person households in the early post-Apartheid period

Wittenberg and Collinson (2007, p.135) have commented that "The national data sets do suggest a veritable explosion in solitary living" for which they find no support in the Agincourt Demographic Surveillance Site. Indeed there seems no plausible "real" mechanism that could have fuelled such a rapid rise. Comparing the estimates from the household surveys to the counts from the 1996 and 2001 censuses[1] (also shown in Figure 1), it seems clear that the most parsimonious reading of this evidence points to a systematic under-sampling of one person (and other small) households in the early OHSs.

Since small households are likely to be materially different from larger households this raises the question whether any reweighting adjustments can be made to the OHSs to bring them into line with the later surveys. Such an adjustment should preferably be made in a way that does not introduce inconsistencies between the person weights and the household weights in the survey. As Branson and Wittenberg (2014) note, the weights for the early OHSs were, in fact, not calibrated to work at both the household and individual level.

Calibration

The theory of calibration is set out in Deville and Särndal (1992). The idea is to minimise a distance function $G(w_k, d_k)$ between the calibrated weights w_k and the design weights d_k subject to a set of external constraints of the sort $\sum_{k=1}^{n} w_k x_k = T_x$ where x_k is some random variable and T_x is the corresponding population total of x. South African datasets are typically calibrated on demographic variables such as province and race-age-gender cells, i.e. the x_k variables are typically dummies specifying membership of particular sub-populations.

[1] To make the comparisons fairer the census counts include only the housing units plus hostel dwellers. We discuss the 1996 census estimate in more detail below.

We do not have access to the original design weights of any of the surveys, so we recalibrate the person weights released with the surveys (see also Branson and Wittenberg 2014). We use the following constraints:

A. Demography
    a. Province totals, as estimated according to the ASSA 2008 model, interpolated to month of the survey
    b. Race-gender-age (in 5 year bands) cells, as estimated according to the ASSA 2008 model, interpolated to month of the survey
B. Constant weights within households
C. Total number of one-person, two-person and three-person households using the 1996 and 2001 census totals as anchors and interpolating for the month of the survey for the periods in between.

We drop the last constraint for OHS 1999 and the Labour Force Surveys. The reason why we still recalibrate those is to ensure that the demography underpinning the series is coherent, as argued for in Branson and Wittenberg (2014).

It is evident that we are calibrating at two levels: on the distribution of individual characteristics (part A) and on the household distribution (part C), while also maintaining a link between person weights and household weights (part B). As such this is an example of "integrated weighting" as discussed by Estevao and Särndal (2003). As noted in that paper, there are two different ways of reconciling the levels of the constraints: we can convert the person-level constraint to a household-level one, or *vice-versa*. To make this more definite consider the two constraints:

$$\sum_{h=1}^{H} w_h z_h = T_z \,, \sum_{i=1}^{n} w_i x_i = T_x$$

where z is a household level variable and x is an individual level one and *H* and *n* are the sample number of households and individuals respectively. We can write this either as

$$\sum_{i=1}^{n} w_i \frac{z_h}{n_{hi}} = T_z \,, \sum_{i=1}^{n} w_i x_i = T_x$$

or

$$\sum_{h=1}^{H} w_h z_h = T_z \,, \sum_{h=1}^{H} w_h t_{xh} = T_x$$

where n_{hi} is the size of household *h* in which individual *i* lives and t_{xh} is the total over x in household *h.* We have presumed that $w_i = w_h$.

In our case we do the calibration at the household level. It turns out that there is an additional choice, because we can write the household constraint either in terms of the total number of one-person, two-person and three-person **households**, or in terms of the

total number of **people** living in one-person, two-person and three-person households. Although the constraints look identical, the underlying calibration turns out to work much better in the latter case.

The calibration estimator that we used is the minimum cross-entropy estimator described in Wittenberg (2010) and Branson and Wittenberg (2014). It is equivalent to the multiplicative calibration estimator described in Deville and Särndal (1992).

In order to weight up the small households one needs to have a reasonable external benchmark to calibrate that distribution to. The 1996 census is the obvious data source. Unfortunately the census household file excludes all hostel dwellers, but hostels are part of the population that the household surveys cover. We can get a count of overall number of hostel dwellers in the 1996 census, but we do not have a sense how many of them live in one-person, two-person or three-person households. To convert the population of hostel dwellers into corresponding household numbers we use the 2001 household distribution among hostel dwellers (taken from the 10% sample, weighted up to the population), e.g. we find that in 2001 43% of all hostel dwellers live in one-person households. We apply that proportion to the 1996 census to estimate that there should be an additional 245 000 one-person households over and above the 1.5 million enumerated in the household file. We make similar adjustments to the counts of two- and three-person households.

Diagnostics

At the minimum the calibrated weights should be able to reproduce the totals to which they have been constrained. The effect of the calibration is evident in Figure 2. We can see the linear interpolation of the total number of households in the period up to October 1998. It is clear that this raises the numbers of both single person households and two person ones. It is interesting to note that three person households are hardly affected. It should be noted that we did not constrain any other part of the distribution, nor did we impose an external constraint on the total number of households.

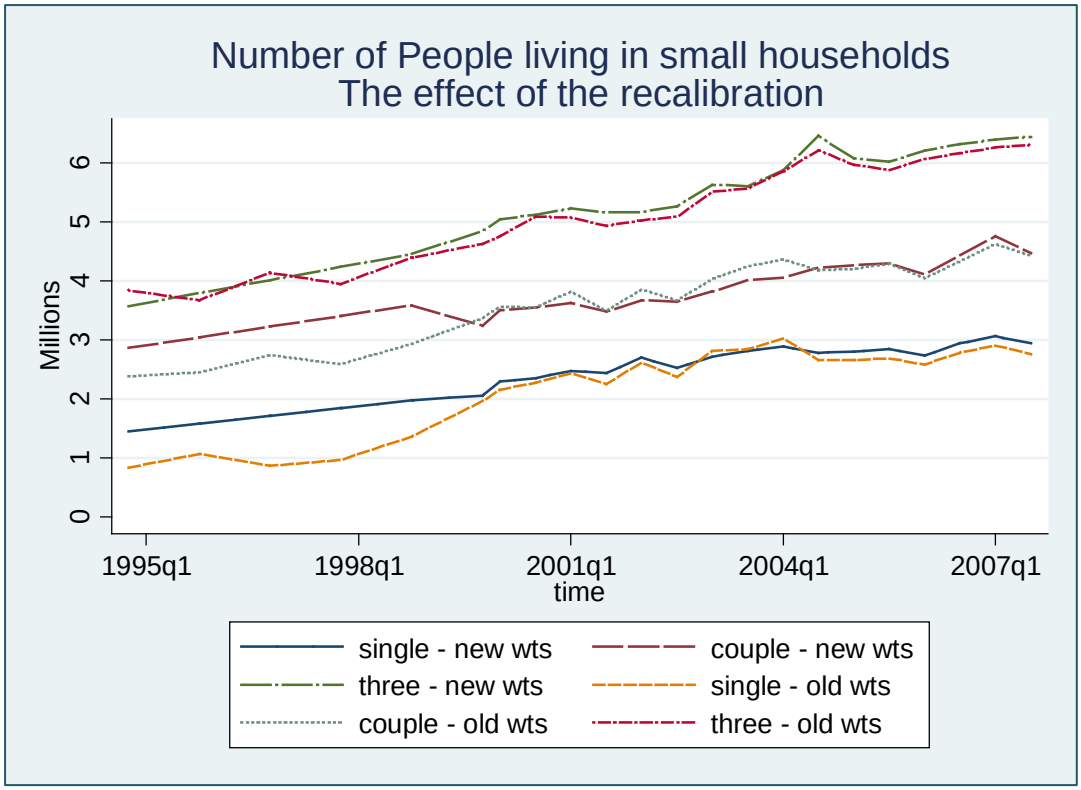

Figure 2 The household size distribution before and after recalibration

The impact on the estimated total number of households is shown in Figure 3. There are three lines in the diagram corresponding to three estimates:

- an estimate calculated with weights recalibrated with all three constraints (A, B and C) for the OHSs up to 1998 and then weights recalibrated to a consistent demography (constraints A and B) for 1999 onwards, marked “D+H” (i.e. demography and household)
- weights recalibrated for the entire period using constraints A and B only, marked “D” (i.e. demography only)
- and the originally released household weights, marked “Old”

The almost linear increase at the beginning of the period is obviously due in part to the linear increases imposed on one-, two- and three-person households. The static pattern exhibited by the “old” (i.e. original) weights in the case of the October Household Surveys is due to the fact that the household weights released with those surveys were not integrated with the person weights.

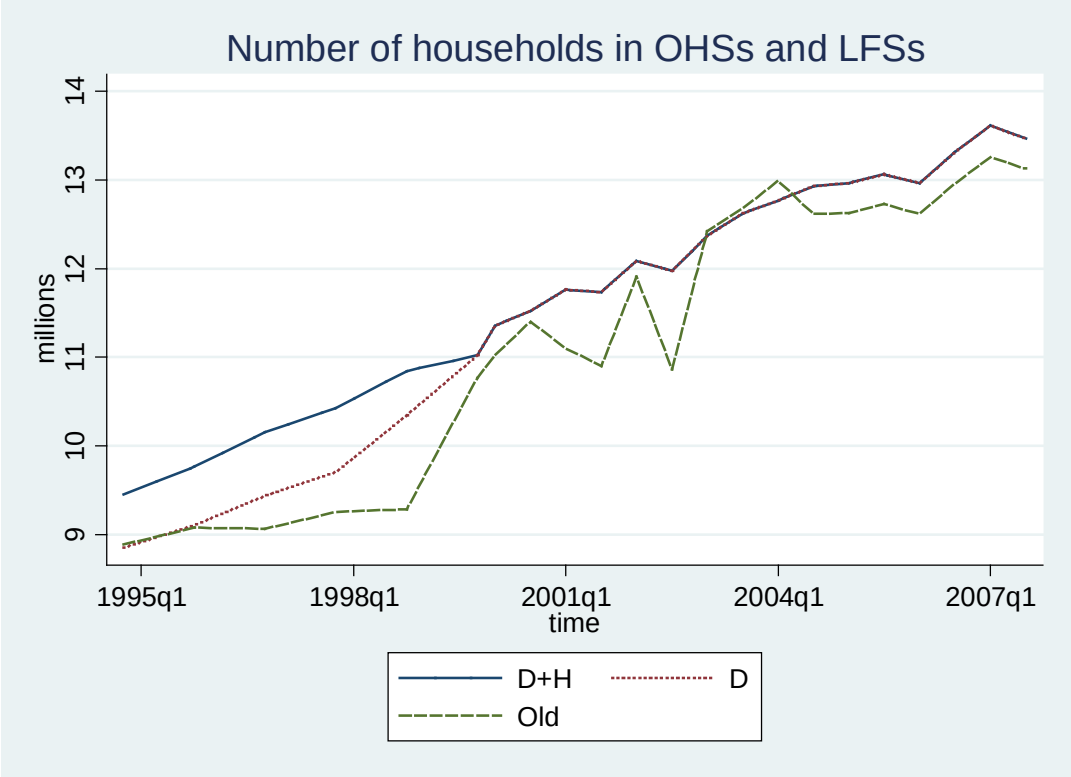

Figure 3 The effect of the recalibration on the estimated total number of households

The overall impact on average household size is shown in Figure 4. The estimate weighting up the small households suggests an almost linear decrease in household size between October 1994 and March 2007, a trend which is not implied by the linear increases we imposed on small households. Obviously this pattern of reduction in household size is more plausible than the one suggested by the original weights.

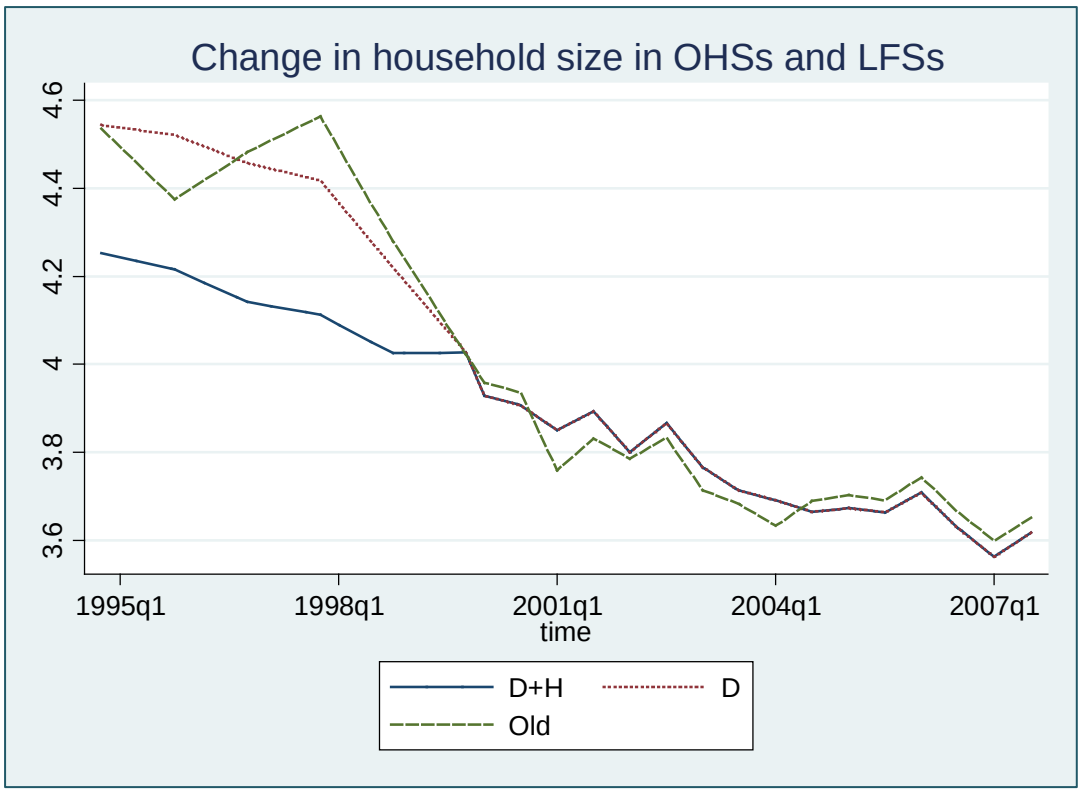

Figure 4 The effect of the recalibration on average household size

Impact on other measurements

It is clear that any measures related to the household size distribution will be affected. It is less clear what impact this will have on other measures extracted from the OHSs. It is important to note that implicit in our procedure is the assumption that the under-sampling of small households did not lead to the complete exclusion of certain subpopulations. We are assuming, for example, that the one person households that we do observe in the samples are reasonably representative of all one-person households, so that we are not exacerbating any coverage errors by weighting up unrepresentative small households. Unfortunately this is not a hypothesis that we can test on the October Household Surveys; the best that we can do is to analyse the impact of the recalibration on a set of other measurements and to judge the plausibility of the resulting trends.

In Figure 5 we have shown the impact of the recalibration on estimated counts of the numbers of informal houses. In Figure 6 we show the corresponding picture for formal housing.

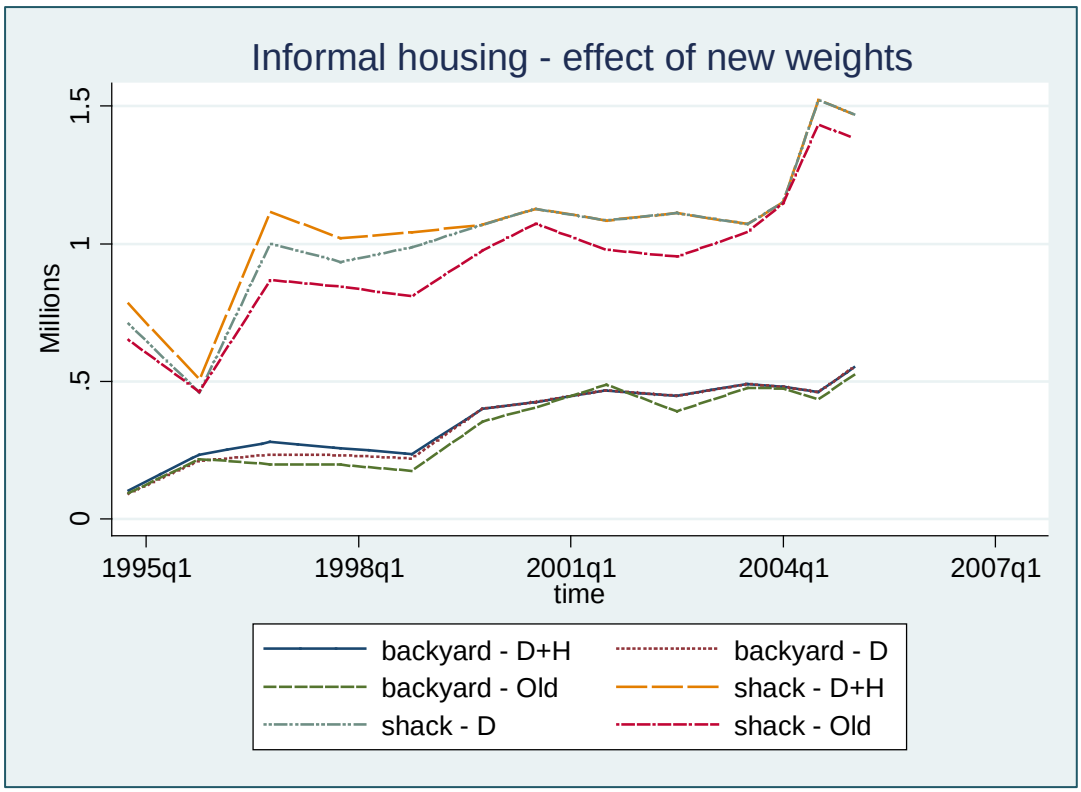

Figure 5 The impact of the weights on measures of informal housing

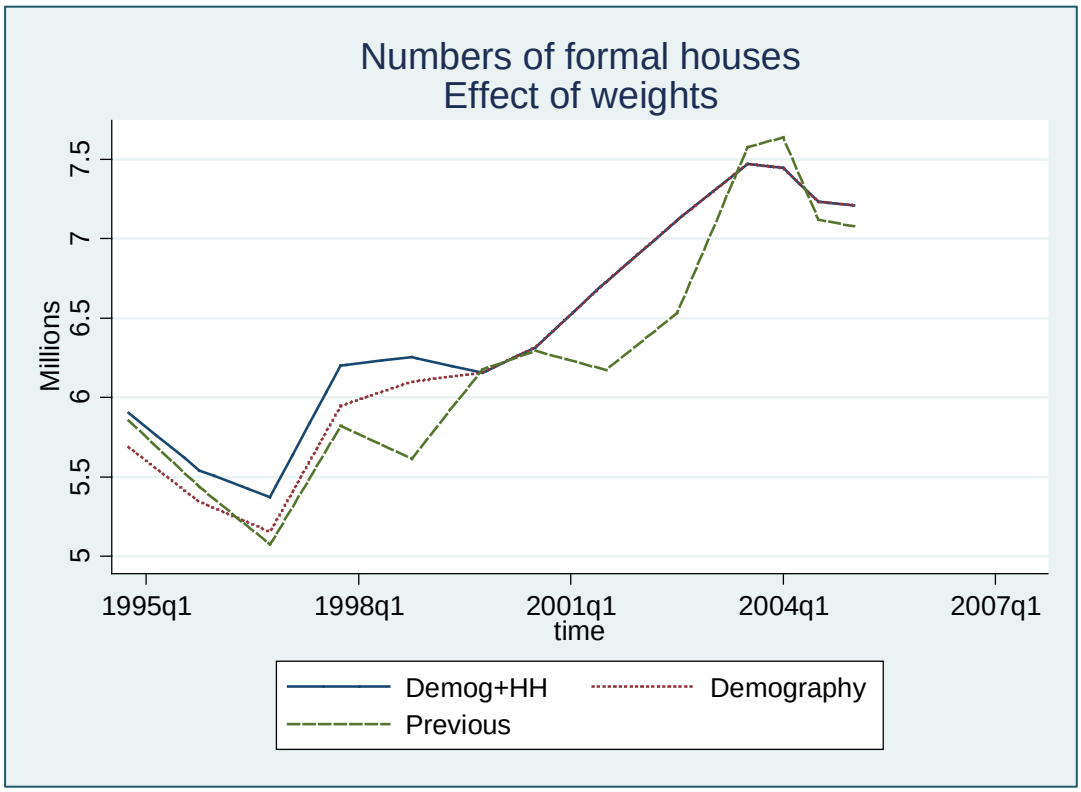

Figure 6 The impact of the weights on counts of formal houses

Several points stand out from these graphs. Firstly, the impact of calibrating to different demographic totals and integrating the person and household weights makes a relatively larger

impact than forcing up the number of small households. Nevertheless weighting up the small households still makes a discernible difference. Secondly, the "backyard shack" series still looks too small in the early OHSs. Clearly the weighting up of small households did not sufficiently weight up this dwelling type. Thirdly, the reweighting does not make the October Household Survey 1995 look any less anomalous (particularly in relation to shack dwellings).

We would anticipate some changes in the distribution of dwelling types and household services. In Figure 7 we show that there is a marked impact on measured employment – but most of the impact again comes from working with a different set of demographic totals. The 2008 ASSA model implied a larger population for the country than the more contemporaneous estimates did.

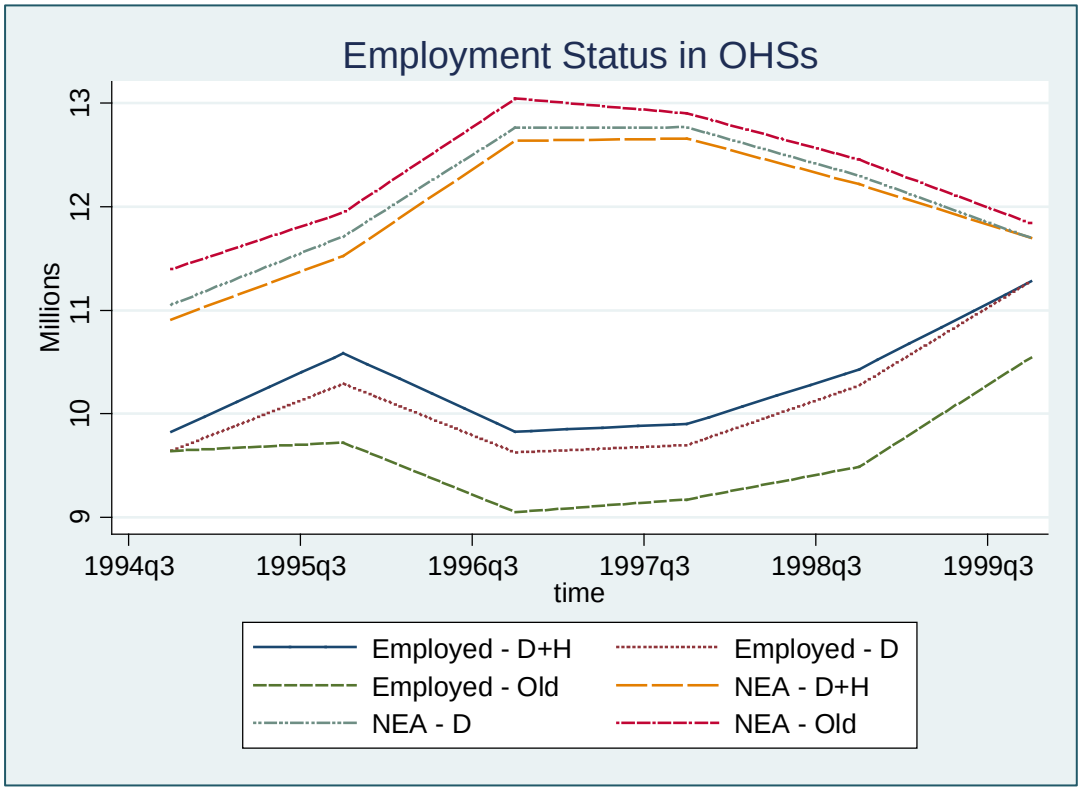

Figure 7 The impact of the recalibration on employment status in the OHSs

The slight upward revision in employment due to the weighting up of small households is of the order of 100 000, i.e. around 1%. That is obviously not massive (it is small when compared to the big jump in the employment numbers between the OHSs and the LFS), but it is of some import. Interestingly, weighting up small households **does** make some difference in two sectors where we think workers might have been missed. Figure 8 shows the picture for domestic workers. Live-in domestics would have been the classic small second household that might have been undersampled by the strategy of only sampling one household at each address. Figure 8 omits 1994, because all domestic workers were captured in the "services" sector in that year. We see that the reweighting increases the point estimate by around 80 000 in most of the early OHSs. Nevertheless the series as a whole jumps around in ways which suggests that the coverage issues are not resolved by the reweighting.

The picture for the mining sector is shown in Figure 9. Again there are sizable adjustments: the point estimate increases by around a third in 1995, 1997 and 1998. Indeed the reweighted series now shows no real break between the OHSs and the LFSs. The figures for some of the other years, in particular 1994 and 1996 remain anomalous. Evidently reweighting is not helpful in contexts where there are no (or insufficient) observations to weight up. Coverage failures can therefore not be "fixed" by after the fact adjustments.

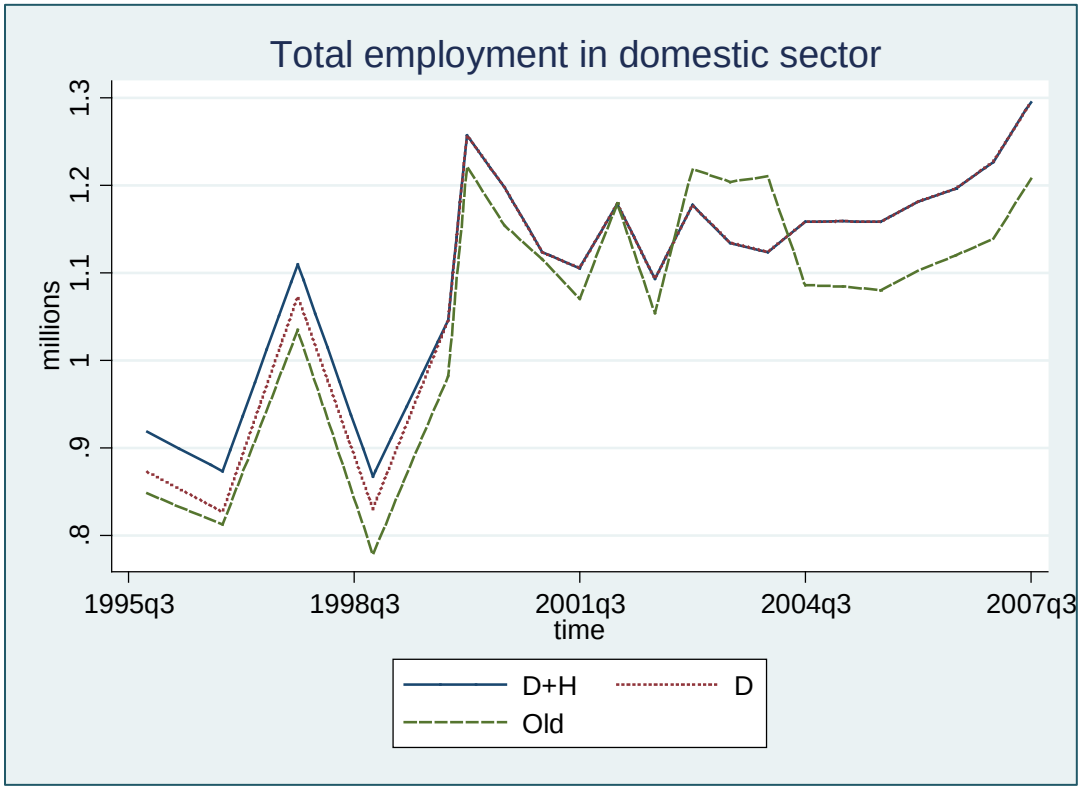

Figure 8 Employment of domestic workers 1995-2007

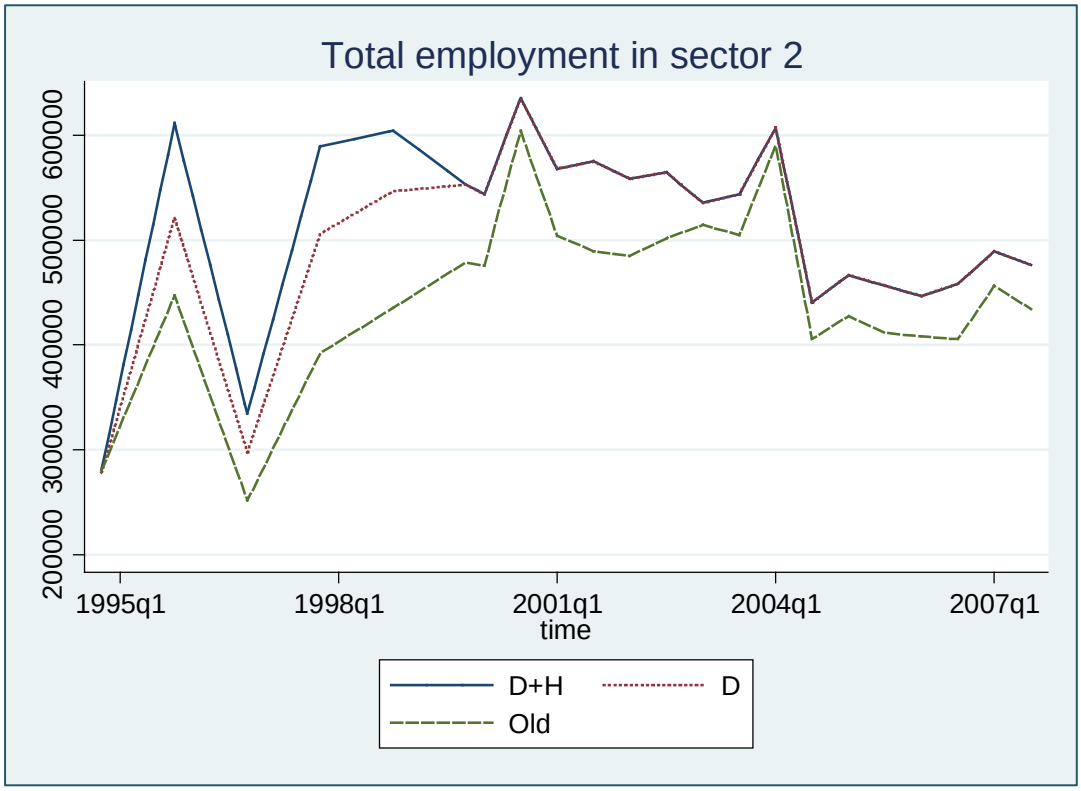

Figure 9 The effects of recalibration on employment estimates in the Mining and Quarrying Sector

Conclusion

Overall the results of the exercise are mixed. The recalibrated weights with household constraints produce time trends on the total numbers of households (Figure 3) and average household size (Figure 4) which are much more plausible than the original ones do. There are also some effects on the counts of formal and informal households and the reweighting produces big increases in the estimated numbers of domestic workers and mine workers, in at least some of the OHSs. Integrating the household and person weights and calibrating to a consistent demography makes a noticeable difference in all cases. Nevertheless it is clear that some of the "missing" data (e.g. backyard shacks pre-1999) remain missing after the recalibration exercise.

What other options for "fixing" the OHSs might there be? Getting access to the original design weights and more information about actual sampling practice would help. Unfortunately some of the key sampling decisions (e.g. which one of multiple households would be interviewed) seem to have been taken in the field. It seems unlikely that one will be able to reconstruct this accurately.

It is therefore likely that we will need to use the information in the existing surveys to gain as much traction on the early post-Apartheid period. Working simultaneously on the distribution of personal characteristics and on those of households is likely to be part of the solution. Our attempts at recalibration show that it is possible to make at least some progress in that direction.

Datasets

DataFirst, Post-Apartheid Labour Market Series [dataset], Version 2.1, Cape Town: DataFirst [producer and distributor], 2013. [zaf-datafirst-palms-1994-2012-v2.1]

Statistics South Africa, October Household Surveys 1994-1999 [datasets], Pretoria: Statistics South Africa [producer], Cape Town: DataFirst [distributor], 2013.

Statistics South Africa, Labour Force Surveys 2000.1 – 2007.2 [datasets], Pretoria: Statistics South Africa [producer], Cape Town: DataFirst [distributor], 2013.

References

Branson, Nicola and Martin Wittenberg (2007), "The measurement of employment status using cohort analysis, 1994-2004", ***South African Journal of Economics*****,** 75(2):313-326.

----- (2014) "Reweighting South African National Household Survey Data to Create a Consistent Series over Time: A Cross-Entropy Estimation Approach", ***South African Journal of Economics***, 82(1):19-38.

Deville, Jean-Claude and Carl-Erik Särndal, (1992), "Calibration Estimators in Survey Sampling," ***Journal of the American Statistical Association***, 87 (418), 376—382.

Estevao, Victor M. and Carl-Erik Särndal, (2003), "A New Perspective on Calibration Estimators", Proceedings of the Survey Research Methods Section, American Statistical Association, pp.1346-1356. Available at https://www.amstat.org/Sections/Srms/Proceedings/y2003/Files/JSM2003-000462.pdf

Kerr, Andrew, and Martin Wittenberg, (2013), "Sampling methodology and field work changes in the October Household Surveys and Labour Force Surveys", DataFirst Technical Paper Number 21, Cape Town: DataFirst, University of Cape Town.

Wittenberg, Martin (2010) "An introduction to maximum entropy and minimum cross-entropy estimation using Stata", ***Stata Journal***, September, 10(3):315-330.

---- (forthcoming), "Data Issues in South Africa", in Haroon Bhorat, Alan Hirsch, Ravi Kanbur and Mthuli Ncube (ed), ***Oxford Companion to the Economics of South Africa***, Oxford: Oxford University Press.

Wittenberg, Martin and Mark Collinson (2007), "Household Transitions in Rural South Africa, 1996-2003", ***Scandinavian Journal of Public Health***, 35 (suppl69): 130-137

About DataFirst

DataFirst is a data service dedicated to making South African and other African survey and administrative microdata available to researchers and policy analysts.

We promote high quality research by providing the essential research infrastructure for discovering and accessing data and by developing skills among prospective users, particularly in South Africa.

We undertake research on the quality and usability of national data and encourage data usage and data sharing.

DataFirst

www.datafirst.uct.ac.za
Level 3, School of Economics Building, Middle Campus, University of Cape Town
Private Bag, Rondebosch 7701, Cape Town, South Africa
Tel: +27 (0)21 650 5708
info@data1st.org / support@data1st.org